Calibrated Peer Reviews in Requirements Engineering Instruction: Application and Experiences

Bastian Tenbergen
Department of Computer Science
State University of New York at Oswego, USA
bastian.tenbergen@oswego.edu

Marian Daun
paluno – The Ruhr Institute for Software Technology
University of Duisburg-Essen, Germany
marian.daun@paluno.uni-due.de

Abstract

Instructing Requirements Engineering (RE) is a challenging task due to the absence of single absolute and correct solutions computer science students so often strive for. Instead, there is often a variety of compromise solutions for each RE problem. Therefore, it is essential that aspiring Software Engineers are exposed to as many solution alternatives as possible to experience the implications of RE decisions. To facilitate this, we propose a learning-by-multiple-examples process, in which we make use of a calibrated peer review grading model for assignments. Paired with a think-pair-share model of semester-long, industry-realistic, project-based low-stakes milestones, we were able to generate a rich collaborative learning atmosphere. In this paper, we report the course design and experiences from the application of calibrated peer reviews in an undergraduate RE course. Qualitative and quantitative application results show that calibrated peer reviews significantly improve students' learning outcomes.

1. Introduction

Requirements Engineering (RE) is one of the most central activities during the development process [1], if for nothing else than it is also one of the earliest. It has been shown over the past decades that high quality requirements are the foundation of high-quality software products and successful development projects [2]. However, studying RE and becoming an effective requirements engineer is a monumental task for learners. On the one hand, they are faced with a vast amount of theory [3]. On the other hand, a substantial amount of experience is required to develop sensitivity to requirements quality, attention to detail, and getting used to the abstract nature of RE [4]. However, students also need to have a sound theoretical basis to apply the principles of RE, such as elicitation, documentation, modeling, and requirements validation. Yet, RE theory instruction is often perceived as boring and unengaging, and like other types of formative instruction [5], often employs rote memorization without building a sound conceptual understanding within the learner.

Thus, in RE education there is a need to (a) for learners to gain as much hand-on experience as possible with different problem situations, while at the same time (b) convey the theoretical basis to understand all relevant aspects of RE. For the former (a), it has been shown that one of the best ways to teach RE in an engaging and motivating way is to make use of non-trivial projects [5] that allow students to explore the problem space in a low-stakes environment (e.g., [6]). In other words: summative instruction [7] allows for creating meaningful learning experiences. For the latter (b), established curricula like SWEBOK [8] or the IREB syllabus [9] exist and define *what* to teach, yet leave it to the instructor to select the best method for *how* to teach it.

To bring (a) and (b) together, a combination of formative and summative instruction argued to be optimal for higher education experiential learning [10], especially in theory-heavy disciplines. We hence propose a hybrid formative/summative learning approach for RE education. The approach consists of three major components:

- *Summative learning:* Realistic, low-stakes case examples [11] are used in combination with theory during lectures and formative assignments.
- *Think-Pair-Share* [12]*:* Students attempt to solve assignment sheets in pairs before presenting their solution to the entire class so all students may see the solution and provide feedback. In this work, we replace the "share" option of Think-Pair-Share with Calibrated Peer Reviews (CPR, see [13]).
- *Calibrated peer review:* Students evaluate, grade, and provide feedback to each other's solutions, leading to increased exposure to solution alternatives, thereby honing experience with the subject matter.

In this paper, we present our approach to integrate CPR in an undergraduate senior-level safety requirements engineering course together with proven experiential learning case studies. To do so, we discuss the background and related work in Section 2 before we present the course design in Section 3 and application of CPR in the course in Section 4. We provide a comparison of assignment and exam grades of our CPR approach to previous non-CPR semesters as well as experiences gained during application in Section 5. Section 6 concludes the paper.

URI: https://hdl.handle.net/10125/79437
978-0-9981331-5-7

2. Related Work

This section reviews challenges and avenues in RE education by overviewing RE topics and giving examples, and discuss CPR [13] as an instructional technique.

2.1. Summative Learning in RE Education

As outlined in Section 1, becoming an effective (i.e., industry-ready) requirements engineer is a monumental task for software engineering (SE) students. Students are essentially asked to learn three things at once: RE theory, including elicitation (e.g., [15]), documentation using natural language (e.g., [16]), modeling languages (e.g., [17]), formal methods (e.g., [18]), and management aspects such as requirements tracing (e.g., [19]) or negotiation (e.g., [20]). Secondly, students are asked to develop a sensitivity for requirements quality, such as completeness and correctness (e.g., [18]), consistency (e.g., [17]), or adequacy (e.g., [21]), and more. Finally, students are asked to learn all these things in a way that makes them effective in industry settings (e.g., [22]), with the client and/or customer in mind (e.g., [23]), and possibly in distributed collaborative settings (e.g., [24]).

Without a doubt, this is too much to master in a single introductory SE course. Therefore, between 2000 [25] and 2012 [26], more and more curricula were adapted and courses have been developed to foster role-specific SE education [27], especially for the role of the requirements engineer [26]. Even early RE education research values summative learning approaches (i.e., learning that allows students to explore the problem space [7], [10]), for example through active student collaboration on projects (e.g., [28]). To this day, project-based (see e.g., [20], [21], [22]) or collaboration-based (see e.g., [23], [28], [29]) instruction is among the favored approaches to instruct RE, either by involving games (e.g., [29]), stakeholders (e.g., [31]), role-playing (e.g., [6], [20]), or case studies (e.g., [5]).

Yet, to become an effective requirements engineer also requires a lot of experience, which other than doing it more often, can also be formed through repeated, low-stakes exposure to examples. To this end, education research favors the combined application of summative learning with formative learning (i.e., learning by remembering theory and concepts, see [7], [10]). Applied to RE, this means students ought to be exposed to theory, followed by formative assignments, and eventually summative projects (like in [29]).

2.2. Calibrated Peer Reviews

CPR [13] is a formative peer-assessment technique, where learners evaluate one another in a systematic way. CPR roughly consists of four phases:

(1) *assignment preparation:* students prepare own solutions to an instructor-assigned problem
(2) *submission and re-distribution of solutions:* anonymized student solutions are distributed to peers
(3) *peer evaluation:* students evaluate other students' solutions against calibrated examples
(4) *feedback collection:* students submit their assessment of others' work along with comments

In addition to the task assignment, the instructor provides a set of example solutions to the students serving as a baseline to help students discriminate solutions of poor, intermediate, or high quality. Calibration examples are either supplied to students during assigning the task (phase 1) or during re-distribution (phase 2) of solutions.

CPR was initially designed as a peer assessment technique for writing assignments [32], [39]. However, it has successfully been applied in higher education at large. For instance, in general science classes [33], more specific in environmental chemistry courses [34], Neuroscience [35], and Engineering [36]. The main advantage of CPR is that students are exposed to more examples of adequate solutions as well as possible pitfalls (so they can avoid them in the future [32], [36]). In addition, their practical skills are increased, especially for non-trivial, non-intuitive topics [34], [35]. Furthermore, the instructors benefit from reduced effort required for grading the work [36], [37], especially in very large courses [39]. Moreover, students have been shown to be more engaged and accept objective criticism more easily when delivered through peers [37]. Disadvantages include increased overhead to create calibration solutions along with collection/redistribution of student solutions, which may be particularly daunting without tool support [38].

Curiously, CPR seemingly has not yet seen widespread adoption in SE education. A notable exception, however, is the very recent work by Aniche et al. in [40]. The authors report on their use of formative peer assessment in a SE course with over 900 student assignment submissions. Results seem to agree with prior work on CPR in that peer assessment reduces workload on instructors while at the same time, yielding a reasonable approximator for grades assigned by instructors. Nevertheless, Aniche et al. found self-assessed grade inflation by about 8-10%, which also agrees with typical variability in peer assessments through CPR.

3. RE Course Design

In this section, we present the course design into which we integrated CPR for assignments. Although the course design was discussed in detail in [29], we present the key elements in the following to provide the reader with a self-contained description. We also place emphasis on aspects that correspond to formative and summative learning as it pertains to the application of CPR to foster comparability with the results from prior work.

3.1. Degree Program & University Setting

The course called "Safety Requirements Engineering" (in the following: RE course) is taught at the State University of New York at Oswego in the Spring semester of each academic year. Housed within the Department of Computer Science, it is a required course for undergraduate students enrolled in the SE baccalaureate program. Students of other department majors may take the course for elective credit, which includes Information Science, Computer Science, and Cognitive Science undergraduate programs as well as graduate programs in Biomedical Health Informatics and Human Computer Interaction. The course's only prerequisite is an introductory SE course, so students have previous exposure to processes, tools, patterns, and topics such as software architecture and software testing. Moreover, most students take three levels of programming courses as well as a computational theory course (covering topics like UML, automata, and well-formedness) before advancing to the RE course in their junior or senior year.

3.2. Topics and Learning Outcomes

The RE course was a key addition to the SE BS degree program to attain ABET[1] accreditation, which was awarded in Summer 2020. To meet accreditation requirements, several topics related to safety and security requirements were added over the design presented in [29], one of which was course work to analyze security risk and their mitigation by eliciting, documenting, and validating security requirements as well as their impact on system safety. Nevertheless, the key focus of the course was to instruct the principles of RE for safety-critical systems. Topics included, but where not limited to requirements elicitation and documentation using natural-language and visual languages; goal- and scenario-oriented RE including misuse cases and attack scenarios; documentation of static-structural, functional, behavioral, and contextual requirements; safety engineering foundations and lifecycle; safety argumentation; safety, hazard, risk, threat, and vulnerability analyses.

Learning outcomes have been formulated for ABET accreditation as follows:

1. Demonstrate in-depth understanding of the different types of requirements and types of requirements artifacts; elicitation and documentation of requirements in various specification formats, throughout several, iterative milestones and at various levels of abstraction.
2. Differentiate requirements that are adequate for the operational purpose of some system from "poor" requirements; conduct relevant analyses to detect and correct defects in requirements impairing the safety, security, and functional adequacy; think abstractly about system functionality and its impact on development.
3. Articulate the (dis-)advantages of solution choices given a problem scope; articulate engineering results to various types of stakeholders.

3.3. Course Design

In this 3-credit course, classes typically meet three times a week for 55 minutes for 15 weeks during the academic semester. Students are expected to invest approximately 10 hours of work outside of class meetings on course assignments and projects. Class meetings are dedicated to the following learning components:

Lectures are the foundational formative learning component in the course. They are used to convey theory, principles, and concepts underlying RE. Support materials consist of slides and reading materials (e.g., excerpts from textbooks, academic articles, and tutorials). Material presentation focuses on concepts and relationships, intertwined with best practices suggestions.

Assignment Sheets are the foundational formative assessment component in the course. Six biweekly assignment sheets are assigned and graded by the instructor. Assignment sheets target specific theory and concepts instructed in the lectures. This follows the "Think-Pair-Share" paradigm [12]: First ("Think"), students are exposed to theory in lectures. Theory is followed up with example problems. Afterwards ("Pair"), students prepare solutions to new problems at the same level of difficulty in teams of two (occasionally three). Students have ten days to complete the assignments. Finally ("Share"), before CPR was added to the course design, solutions were discussed in class. Since in RE, there is rarely a single, optimal solution to a problem, assignment sheet discussions relied heavily on students showing their solutions and discussing different approaches, their advantages, and disadvantages. This sometimes took several class meetings. During this phase, the instructor graded all solutions, infusing the discussion with solution tips.

Industry-realistic ***case example projects*** make up the summative learning and assessment part of the RE course. Their application has been documented in detail in [29] and aims at providing hands-on experience with diagram notations, relationships between artifacts, purpose and meaning of concepts discussed in class, etc. Realistic specifications such as an airborne collision avoidance system, automotive driver assistance

[1] ABET is a non-governmental organization accrediting engineering degree programs in the US, see http://www.abet.org

systems, an automotive remote key locking system, or autonomous industrial transportation robots are used for this purpose. In three to four comprehensive milestones, students are asked to produce a correct and internally consistent and ambiguity-free specification of traceable requirements, including hazard, threat, and risk analyses as well as mitigating requirements. At several points during the semester, students share their preliminary results for feedback and a partial instructor-assigned grade. Project milestones were prepared in teams of four students (i.e., two assignment sheet teams pair up).

A ***midterm and a final exam*** with questions and tasks focused on documentation and analysis techniques as well as theoretical concepts as a measure of understanding of relationships between concepts and techniques rounded assessment of learning objectives in this course. The midterm is mainly formative in nature, as it contains problems like those encountered in the assignment sheets. The final exam is mainly summative in nature as it required students to produce a very small requirements specification for a new, but thematically related case example system.

3.4. Past Observations regarding the Learning Experience

Experiences made in the original RE course (i.e., without application of CPR) were initially reported in [29] in large detail, including quantitative results. In the following, we provide a brief overview to lay a foundation for comparison after the application of CPR:

Lively discussion with strong focus on practicality. The course traditionally emphasizes lively discussion (to the point of open disagreement and a high degree of dynamicity). This would often take the form of students volunteering to show their solutions, arguing advantages and disadvantages of a solution choice amongst each other, and drawing examples on the white board to make their points. Often, the instructor acted as a moderator of the discussion, rather than the "grading authority." Many students sought further feedback outside of class by meeting with the instructor. By student feedback, we gauge this mode of interaction to be the key driver to fostered RE knowledge in students.

Teamwork and eagerness to engage in class proceedings. The lively discussion culture that naturally evolved in the course resulted in a very high degree of engagement in almost all students. With only few and occasional exceptions, all students regularly attended class meetings and team meetings outside of class, made themselves available for teamwork, and engaged in class discussions.

Need for dynamic adjustment of class content, depending on semester progress. Lively discussions were sometimes hard to contain by the instructor, thereby resulting in class meetings in some instances taking a turn towards discussion, rather than theory instruction. This often required us to adapt the semester plan dynamically according to how the semester progressed, sometimes requiring us to combine or reorganize lectures to ensure sufficient instruction before assignment sheets were due. While an effort was made to maintain all topics and learning outcomes, this occasionally resulting in assigning mandatory reading to students.

Reduction in students' preoccupation with solutions desired by the instructor. Typically, students are preoccupied with instructor-desired solutions in the beginning of the semester, often asking "How do you want this to be done?" or "Is this what you wanted?". As the semester progressed, almost all student teams seemed to naturally transition to confidently presenting and defending their solution. While some students struggled with this, others adjusted so well that they began to occasionally challenge others' ideas, including the instructor's (which the instructor encouraged).

Steep learning curve regarding safety, and requirements quality. While students' artifacts were typically vague, abstract, and superficial at first, repeated exposure to examples, assignment sheet solutions, case example specifications, and the ability to receive low-stakes feedback on preliminary case study results eventually allowed most students to develop sensitivity for insufficient detail and conflicting information within their solution (e.g., ambiguous requirements). For example, students developed relatively good skills in refining high-level goals into concrete system functions. On the other hand, this seemed to be a bit more difficult with safety mitigations, which often took the form of merely stating the opposite of a hazard. For example, the hazard "airbag deploys too early" would be refined into a safety-goal "make sure airbag doesn't deploy too early". Developing functionality that would instead minimize the risk of this happening was less obvious and came less naturally to many students.

4. Application Calibrated Peer Reviews

In this section, we present the specific changes to the course design from Section 3.3 to accommodate CPR. We begin by discussing the motivation of these changes.

4.1. Motivation

Since RE is a socio-technical process [41], it was unsurprising to us that subjectively most successful quality of the RE course to date were in-class discussions about solution alternatives. In fact, from the instructors' points of view, this was one of the most pleasurable aspects in all semesters from Spring 2017 to Spring 2019 [29]. However, in Spring 2020, due to the COVID-19

pandemic, this quality was suddenly lost when all in-person instruction was suspended and shifted to an all-online mode by gubernatorial order [42].

The RE course seamlessly transitioned to synchronous online class meetings using video conferencing and (to the instructor's admiration) all students increased their effort to successfully complete the RE course (a certain "let's get through this together" atmosphere established itself in the course). Yet, the nature of class presentations and discussions moved away from agile and dynamic interactions to merely presenting progress reports. Before, students from other teams would spontaneously get up from their seat, take a dry erase pen, and contrast their work on the whiteboard with the presenting team's work. Albeit digital whiteboard websites were used during online class meetings, their use with mouse or touchpad allows at best rudimentarily demonstration of concepts, and mainly by the instructor.

Students picked up on this change. In free-text answers during the annual post-semester course evaluations, students appreciated all course members' attempts to "try and make things seem normal." However, many lamented the noticeable decrease in discussions, fewer opportunities to review others' work, and much less exposure to examples. This loss particularly extended to the formative aspects of the course. The summative project milestones allowed showing and receiving feedback on preliminary solutions for the purpose of augmenting them before submission and grading, whereas the formative assignment sheets did not.

We observed a similar trend in a "sister" course on software quality assurance during the following Fall 2020 semester [13]. This course is structured very similarly to the RE course (and in fact, is considered its "counterpart" in the SE BS curriculum) and was offered in a HyFlex format due to the COVID-19 pandemic. Yet, remotely participating students, regardless of synchronous or asynchronous reported much less engagement with other students' examples.

4.2. Implementation

To alleviate these limitations brought forward due to synchronous online instruction, in the Spring 2021 semester, we modified the RE course to maximize student interaction and exposure to examples. Since lectures and project milestone discussions seamlessly transitioned to synchronous online class meetings, we specifically targeted the formative assignment sheets.

Since peer assessment is a suitable means to increase students' exposure to alternative solutions while at the same time stimulating critical introspection about their own work [40], we decided to implement CPR as a mode of assessing the formative assignment sheets. Fig. 1 shows the procedure in which we applied CPR from the instructor perspective (i.e., rounded-corner rectangles depict activities carried out by the instructor).

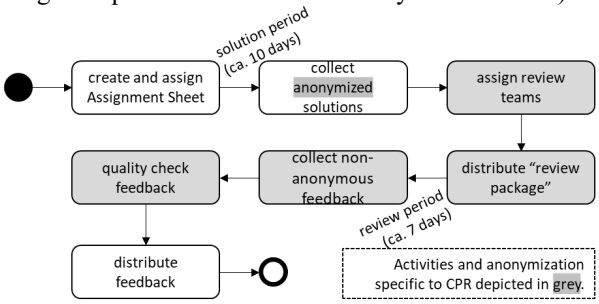


Fig. 1 Procedure of Implementing CPR in Assignment Sheets

Ordinarily (see Section 3.3), the instructor creates and distributes the assignment sheets and collects non-anonymous student solutions after ca. 10 days of preparation period. The university's online learning management platform was used for distribution and collection. After that, the instructor grades the assignment sheets and sends grades and feedback to the students. In Fig. 1, these activities are shown in white.

For CPR, some additional work overhead is required, depicted in grey in Fig. 1. This is as follows:

Anonymization. Since peer assessment in CPR produces a partial grade for the students, US federal regulation requires protection of students' identity and disclosure of graded work [43]. Therefore, assignment sheet teams (usually two students) were asked to select an anonymous, yet specific-to-them team name. Team names were collected during team formation and retained until the end of the semester. Students were asked to pick a name consisting of letters and numbers.

Assign Review Teams. After the solution period of ca. 10 days was over, the instructor collected all student solutions digitally as a PDF file through the online campus system. The instructor would then quasi-randomly assign review teams to other teams' solutions, ensuring that no team would review their own solution. In doing so, an effort was made to balance the review assignments such that each solution received the same number of reviews and each team would review the same number of review assignments. In Spring 2021, this worked out to be exactly three reviews per solution and per team for each assignment sheet. Had this not been the case, the compensation strategy was to ensure equal number of reviews per solution, even if that meant occasionally assigning one extra review to individual teams, making sure that this is then balanced across weeks (i.e., Team "Lo97Ro01" reviews four solutions this week, but only three solutions next week). Furthermore, an effort was made to change reviewer-solution assignments across weeks, thereby ensuring that no solution is unreasonably often reviewed by the same reviewer team. Yet, it was unavoidable to eventually assign a solution from a

previously reviewed team to the same review team again. Table 1 shows the review assignments (and grades) for Assignment Sheet 4 in Spring 2021 as example how anonymous review team assignments were facilitated.

Distribution of "Review Package". Once reviewer assignments were complete, review teams were distributed a "review package." The review package contained the following:

1. Solutions to be reviewed that were produced by the respective other teams;
2. Calibrated solutions of "poor", "medium", and "high" quality;
3. Review instructions detailing common pitfalls and critical success factors to help reviewers differentiate "poor" from "high" quality solutions;
4. A template to document feedback and grades.

While (1) was emailed to each team containing only the assigned solutions, (2-4) were the same for all teams and distributed via the online learning platform.

The "calibrated solutions of 'poor', 'medium', and 'high' quality" (2) were anonymized student solutions taken from previous semesters if such solutions were available. Else, the instructor produced a calibration solution that, given the grading scheme outlined in the review instructions (3), would attain <70% of available points (i.e., constitute a "poor" quality solution), 70%-90% of available points ("medium" quality), or >90% of available points ("high" quality). Calibrations contained example point deductions (or reversely, awarded points) along with brief justifications why points were deducted or awarded. For modeling exercises, the calibration would also contain an instructor-created solution as an example of one minimally acceptable full-score solution.

Calibration examples (2) and an example of review instructions (3) with the grading template (4) is available in the paper supplement[2]. A complete list of course material is available from the first author upon request.

Collect Non-Anonymous Feedback. After a review period of about seven days, feedback and peer-assigned grades were collected via the online learning platform. This time, to make it easier for the instructor to collate feedback for teams, reviewer teams were asked to submit feedback as a non-PDF, editable file (MS Word was preferred). For each team, their respective feedback was then copied into a file for subsequent distribution and assigned grades were collected as shown in Table 1. Albeit it matters not to the implementation of CPR, at this point it should be noted that in the RE course, assignment sheets and project milestones are graded out of 15 points.

Quality Check Feedback. Albeit the instructor routinely quality checks any feedback they give to students, this step received particular attention during CPR. This took two forms: firstly, we checked feedback against obvious unacceptable content, such as empty justifications or profanity. Such feedback was to be filtered out (but not a single occurrence was found during the semester). Secondly, we checked that students made an adequate effort to review other teams and did not assign a default (usually maximum) grade or inexplicably low grades. Table 1 shows an example in team "Br00Mi99", who only assigned the maximum 15 points for all teams. Comparing their feedback against other teams' feedback of the same submission, however, shows that these relatively lenient grades are acceptable as team "Br00Mi99" was assigned three high quality solutions by chance.

Similarly, Table 1 shows that team "An98Wi00" assigned a very low grade (8.5, which is 56.6%) to team "Po94Ch97", while other teams were more lenient with their reviews of team "Po94Ch97". Compared to other feedback issued by the same team, it turned out that one instruction (i.e., "reading directions must be part of class diagram association labels") was interpreted particularly harshly by this team (we will review experiences in more detail in Section 5).

Distribute Feedback. After feedback quality checking completed, we distributed feedback to the recipient teams via email.

Table 1 Example of Peer Review Assessments

Team Name	An98Wi00	Br00Mi99	Br99Da99	Ch99Sa00	Do99	Na99Ma96	Pa99Mi99	Po94Ch97	Avg Grade RECEIVED
An98Wi00		15			15		14		14.67
Br00Mi99	9		14.25		12.5				11.92
Br99Da99						14.25	10.5	13	12.58
Ch99Sa00		15	10.5			14.5			13.33
Do99			13	10		11.5			11.50
Na99Ma96	15			12.5				14	13.83
Pa99Mi99				13.5	15			13.75	14.08
Po94Ch97	8.5	15					13		12.17
Avg Grade GIVEN	10.83	15.00	12.58	12.00	14.17	13.42	12.50	13.58	

[2] Online supplement is available at: http://doi.org/10.6084/m9.figshare.14718831

5. Grade Comparison and Experiences

In this section, we present quantitative findings regarding assignment sheet performance (Section 5.1) and formative learning as assessed through exams (Section 5.2). Section 5.3 discusses qualitative findings regarding students' and instructors' experiences.

5.1. Comparison of Assignment Sheet Grades

Figure 2 compares the average assignment sheet performance for each of the six assignment sheets in the 2019, 2020, and 2021 offerings of the RE course. To facilitate comparison, the assignment sheets were the same across all three semesters. Topics included introduction to RE, elicitation, goal and scenario modeling, as well as data, functional, and behavioral requirements. The column "Final" in Fig. 2 represents the combined average across all six assignment sheets for each year.

As can be seen from Fig. 2, apart from assignment sheet 3 (goal and scenario modeling), performance is consistently at or above the performance from both previous years. Similarly, the combined average is roughly nine percentage points above the previous best performance in 2020. Since the assignment sheets and mode of solution preparation were the same in all three semesters, the difference in score must be attributed to the mode of assessment. In 2019 and 2020, assessment was done by the instructor, but in 2021 was done through CPR. It must be noted that the grading rubric (i.e., what to deduct or award points for) was the same each year.

The results seem to indicate that CPR yields on average to a higher assigned score for a given solution. To test if this difference is significant, we statistically compared the combined average of 2021 against those from the previous two years. Results of three corresponding T-Tests are shown in Table 2. T-Test type was determined by verifying or rejecting equality of variances through a pre-hoc F-Test and post-hoc statistical power was computed to determine the size of the effect.

Table 2 shows that the difference between 2019 and 2020 is not significant. However, the difference between 2021 and both previous years is significant ($p < 0.05$). This means that peer assessment through CPR in 2021 lead to significantly higher assignment sheet grades compared to instructor-assessed grades. This agrees with the results reported by Aniche et al. [40], who also found grade inflation by about 8-10%.

A caveat to this result is the unequal sample size between 2019 and both following years as well as the overall low sample size in 2020 and 2021. Even though post-hoc power analyses determined a large effect size, a small likelihood of false positive results remains [44].

Table 2 Statistical Comparison of Assignment Sheet Performance (averages across all 6 assignment sheets)

	2021	**2020**	**2019**
Mean	89.23%	80.45%	74.37%
Variance	5.43%	18.59%	16.00%
Sample Size	16	16	32

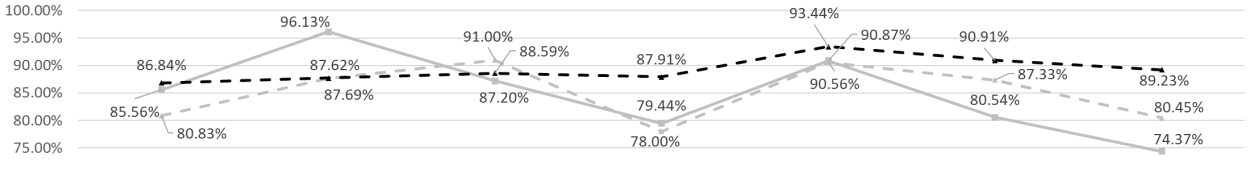
Student's T	**2019 vs 2020**	
dF	26	
F	1.3939	(unequal variances)
p	0.4459	(not significant)
Cohen's d	0.9501	(large effect)

Student's T	**2019 vs 2021**	
dF	46	
F	0.1188	(equal variances)
p	**0.0139**	**(significant)**
Cohen's d	>1.001	(large effect)

Student's T	**2020 vs 2021**	
dF	30	
F	0.0852	(equal variances)
p	**0.0447**	**(significant)**
Cohen's d	>1.001	(large effect)

5.2. Comparison of Exam Grades

In Section 5.1, we established CPR leads to higher assignment score grades. Yet, since this is merely an effect of the mode of assessment of assignment sheets, we need to investigate if CPR influences learning outcomes. In the RE course, the main mode of learning outcome assessment is through the instructor-assessed exams. Exam result averages are shown and compared in Figure 3.



	Sheet 1	Sheet 2	Sheet 3	Sheet 4	Sheet 5	Sheet 6	Final
2019	85.56%	96.13%	87.20%	79.44%	90.87%	80.54%	74.37%
2020	80.83%	87.62%	91.00%	78.00%	90.56%	87.33%	80.45%
2021	86.84%	87.69%	88.59%	87.91%	93.44%	90.91%	89.23%

Fig. 2 Comparison of Students' Assignment Sheet Performance

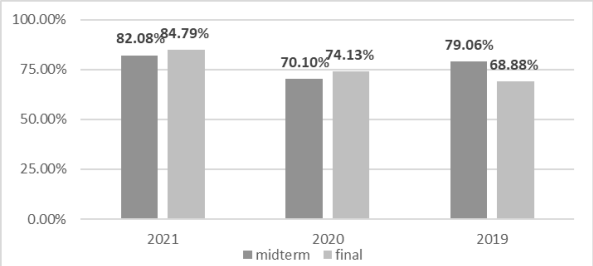


Fig. 3 Comparison of Students' Exam Performance Semesters

As can be seen, just like in assignment sheets, students achieved a higher score in both the midterm and final exams in the 2021 offering of the RE course, compared to the previous years' exams. In fact, in 2021, students performed consistently 10 percentage points better than in previous years. Table 3 shows the combined averages for the exams across all three years.

In all three years, the exams were graded by the instructor using the same grading rubric. Therefore, differences in exam performance can be attributed to the formative learning throughout the semester. Since formative learning made extensive use of CPR in 2021, we hypothesize that CPR increased the exam grades and hence, learning outcomes. To investigate this claim, we conducted T-Tests of the exam averages (again, using F-Tests to check variance equality and using Cohen's d to determine post-hoc effect size).

T-Test results are also included in Table 3 and show that the difference between 2020 and 2019 is not significant ($p > 0.05$). Exam scores are significantly higher in 2021 than in 2019 ($p = 0.0458$). Compared to 2020, the exam scores in 2021 are higher, yet not significant ($p = 0.0502$, we strictly adopt 5% as the significance threshold). We believe that similarly to the results in Section 5.1, unequal and overall low sample size in 2021 may be the culprit for both T-Test results (i.e., 2021 vs. 2019 and 2021 vs. 2020) edging at the significance level of 5%. We presume that comparing the results of more than 16 students would lead to more clear results. Yet power analyses determined a large effect size, thereby suggesting that the increase in exam score is nonetheless be related to the application of CPR. Another caveat is the mode of instruction (online vs. face to face), which, was likely remedied by students' efforts to "cram" for exams. Yet, this factor likely increased the effect CPR had because study material would make use of others' solution provided through CPR. Therefore, we confidently conclude from these results that CPR yields a significant increase in students' learning outcomes, nonetheless.

Table 3 Statistical Comparison of Average Exam Performance

	2021	2020	2019
Mean	83.44%	72.12%	73.97%
Variance	12.32%	22.74%	19.68%
Sample Size	16	16	32

Student's T	**2019 vs 2020**	
dF	26	
F	1.3794	(unequal variances)
p	0.3945	(not significant)
Cohen's d	0.9509	(large effect)

Student's T	**2019 vs 2021**	
dF	46	
F	0.4049	(equal variances)
p	**0.0458**	**(significant)**
Cohen's d	0.9857	(large effect)

Student's T	**2020 vs 2021**	
dF	30	
F	0.2936	(equal variances)
p	0.0502	(not significant)
Cohen's d	0.9644	(large effect)

5.4. Student Reactions and Instructor Experiences

Qualitative experiences from application of CPR can be divided into those reported by the students and observations by the instructor, which we will outline in the following.

Student Experiences. Since this was the first time the instructor applied CPR in a course like this, we inquired for students' honest feedback roughly on a weekly basis, asking for ways in which CPR was successfully applied or aspects that require improvement. Initially, students were rather indifferent about CPR. Some viewed it in the early stages of the course as "yet another thing to do". This attitude shifted dramatically with the second assignment sheet, which asked students to elicit requirements from an interview protocol. Since this was a rich resource, students quickly grew to appreciate being able to contrast their findings with others'.

An issue that occurred twice through the semester involved "outlier" grades. In two occurrences, a reviewer team assigned considerably lower grades to a solution than other reviewer teams (see Section 4.2 and Table 1 for one of the two instances as an example). During the "quality check" phase (see Fig. 1), the instructor took note of the outliers in both cases. We contemplated discarding the feedback and grade, yet it was decided that "counter examples" of appropriate feedback are also valuable. So, when the students receiving the low score complained to the instructor with anxiety over a grade, it was pointed out that (a) this result is an outlier that only minimally impacts the grade (if at all), that (b) they should focus on the written feedback instead of the grade and decide if it is valid, and (c) if the feedback is not valuable to them, take this occasion as an example of what poor quality feedback means for the learner. The instructor had previously already removed outlier grades and

discussed grading guidelines with the respective reviewer team.

Instructor Experiences. The issue of outlier or improper feedback appears dramatic to the students when it happens. Even for the instructor, the prospect of this issue caused some initial anxiety regarding how this would impair the learning experience for the students. Yet, from the perspective of the instructor, this turned out to be a negligible problem. It only occurred twice (out of 8 teams x 3 reviews x 6 sheets = 144 peer reviews) and was very easily and amicably rectified with all involved parties. Other pre-hoc instructor worries included profanity in reviews, empty review templates, or other unforeseeable student behavior undermining the CPR process. To the relief of the instructor and to the credit of his students, none of these issues occurred at all.

We did observe positively skewed peer assigned grades (see Section 5.1). Especially teams, who themselves produce mediocre or low-quality solutions tend to overly inflate grades assigned to others. Conversely, high performing teams hesitate to deduct points from other teams, only deducing quarter or half points even for severe mistakes. As the semester progressed and students gained more experience with grading, grade distribution grew similar to instructor-assigned grades.

Finally, the chief complaint from the instructor perspective is the time it takes to collect, assign, and distribute solutions in the CPR process (see Fig. 1). While the instructor was hopeful to reduce the time needed to engage with assignment sheets, we found ourselves spending *more* time with managing CPR than the time it took to grade assignment sheets in previous years. The process was managed entirely manually and, for a very large course (like in [40]), would likely not have been feasible. This means that tool support is essential. Given proper tool support and once calibration examples are available, we anticipate instructors to save considerable amount of time, especially in large courses.

We conclude that CPR is an excellent mode of formative assessment in RE instruction. During the COVID-19 pandemic, solution exposure and peer-feedback were difficult to implement and much less effective in online setting as students' level of engagement with the class is much lower than in in-person settings (cf. [13]) In consequence, formative aspects of theory instruction may be less effective, diminishing students' long-term retention of material [5]. Our results show that CPR significantly improves this issue and students' learning outcomes. We will continue to apply CPR (with tool support) in future versions of this course.

6. Conclusion & Outlook

In this paper, we have presented a requirements engineering (RE) course design combining experiential learning with formative peer assessment. Specifically, we propose the use of Calibrated Peer Reviews (CPR) as a mode of instruction for assignment sheets to expose students to solution alternatives and engage in critical reflection about RE theory. We have outlined the course design and application of CPR in sufficient detail to invite others to replicate our results. We have reported on qualitative and quantitative results of CPR application.

Results confirm previous work which found grade inflation by about 10% through peer assessment. Results also show that applying CPR significantly improves students RE learning outcomes. Qualitative experiences show that the overhead on the instructor on providing calibration examples and managing the anonymized peer evaluation process is substantial and outweighs the time saved in grading, even in small class sizes. Yet, increased exposure to alternative solutions is a rich, fruitful experience for the learner, both as self-reported by the students and as observed by the instructor. Future work is concerned with exploring tool support for CPR and gathering additional evidence to the effectiveness of CPR.

References

[1] M. Ochodek, S. Kopczńska, "Perceived Importance of Agile Requirements Engineering Practices – A Survey." J Sys & Soft 143, 2018, pp. 29-43.

[2] A. Terry Bahill, S. Henderson, "Requirements development, verification, and validation exhibited in famous failures." J Int. Council on Sys Eng 8(1), 2005, pp. 1-14.

[3] R. Memon, R. Ahmad, S. Salim, "Problems in Requirements Engineering Education: A Survey." Proc. 8th Int. Conf. Frientiers of Info Techn 2010, pp. 1-6.

[4] G. Regev, D. Gause, A. Wegmann, "Experiential Learning Approach for Requirements Engineering Education." Requirements Eng 14, 269 (2009).

[5] M. Daun, A., Salmon, B. Tenbergen, T. Weyer, K. Pohl, "Industrial case studies in graduate requirements engineering courses: The impact on student motivation." Proc. IEEE 27th Conf Soft Eng Edu & Train, 2014.

[6] R. Berntsson Svensson, B. Regnell, "Is role playing in Requirements Engineering Education increasing learning outcome?" Requirements Eng 22, 2017, pp. 475-489.

[7] W. Harlen, M. James, "Assessment and Learning: Differences and Relationships between Formative and Summative Assessment." Assessment in Edu: Principles, Policy & Practice 4(3), 1997, pp. 356-379.

[8] P. Bourque, R. Fairley, "Guide to the Software Engineering Body of Knowledge (SWEBOK®)", Version 3.0. IEEE Computer Society Press, 2014.

[9] IREB, "Certified professional for requirements engineering foundation level syllabus v3.0.1," Int. Requirements Engineering Board e.V., Tech. Rep., October 2020.

[10] A. Man Sze Lau, "Formative good, summative bad? – A Review of the Dichotomy in Assessment Literature." J Further and Higher Edu 40(4), 2016, pp. 509-525.

[11] K. Garg, V. Varma, "A Study of the Effectiveness of Case Study Approach in Software Engineering Education", Proc. 20th Conf Soft Eng Edu & Train, 2007.

[12] M. Kaddoura, "Think pair share: a teaching learning strategy to enhance students' critical thinking." Edu Research Quarterly, 36 (4) (2013), pp. 3-24.
[13] T. Brown, B. Tenbergen, "Teaching Software Quality Assurance (SQA) During COVID-19 using the HyFlex Approach -- Course Design, Results, and Experiences." In Proceedings of the Annual Symposium of the American Society of Engineering Education, 2021
[14] R. Robinson, "Calibrated Peer Review ™ An Application to Increase Student Reading & Writing Skills. The American Biology Teacher 63(7), 2001.
[15] M. Bano, D. Zowghi, A. Ferrari, P. Spoletini, B. Donati, "Learning from Mistakes: An Empirical Study of Elicitation Interviews Performed by Novices." Proc. 26th IEEE Int. Requirements Eng Conf, 2018.
[16] D. Jagielska, P. Wernick, M. Wood, S.Bennett, "How natural is natural language? How computer science students write use cases?" Proc. 21st ACM SIGPLAN Symp OO Prog Sys, Lang, & Applications, 2006.
[17] K. Sikkel, M. Daneva, "Teaching Consistency of UML Specifications." Proc. 5th Int. WS Requirements Eng Edu & Train 2005.
[18] B. Westphal, "An undergraduate requirements engineering curriculum with formal methods." Proc. 8th WS Int WS Requirements Eng Edu & Train, 2018.
[19] O. Gotel, S. Morris, "Case-based stories for traceability education and training," Proc. 7th Int. WS Requirements Eng Edu & Train, 2012.
[20] B. Al-Ani, N. Yusop, "Role-playing, group work and other ambitious teaching methods in a large requirements engineering course," Proc. 11th IEEE Int. Conf. & WS Eng Computer-Based Systems, 2004.
[21] R. Noel, R. Munoz, C. Becerra, and R. Villarroel. Developing competencies for software requirements analysis through project based learning. Proc. 35th Int. Conf. Chilean Computer Science Society, 2017.
[22] J. Fernandes, R. Machado, S. Seidman, "A Requirements Engineering and Management Training Course for Software Development Professionals." Proc. 22nd Conf. Soft Eng Edu & Train, 2009.
[23] D. Suri and J. Gassert, "Gathering project requirements: A collaborative and interdisciplinary experience." Proc. American Soc Eng Edu Ann Conf., 2005.
[24] D. Marutschke, V. Kryssanov, P. Brockmann, "Teaching distributed requirements engineering: Simulation of an offshoring project with geographically separated teams." Proc. IEEE 32nd Conf Soft Eng Edu & Train, 2020.
[25] Shaw, M., Software Engineering Education: A Roadmap. Proc. Future of Soft Eng, 2000, pp. 371-380.
[26] M. Daun, A. Grubb, B. Tenbergen, "A Survey of Instructional Approaches in the Requirements Engineering Education Literature." Proc. 29th Intl. Conf Req Eng 2021.
[27] N. Mead, "Software Engineering Education: How Far We've Come and how Far We Have to Go." J Sys Soft 82 (2009), pp. 571-575.
[28] D. Rosca, "An Active/Collaborative Approach in Teaching Requirements Engineering." Proc. 30th Ann Frontiers in Edu Conf, 2000.
[29] B. Tenbergen, M. Daun, "Industry projects in requirements engineering education: Application in a university course in the us and comparison with germany," In Proc. Hawai'i Intl. Conf. Sys Sci, 2019.
[30] M. Alexander, J. Beatty, "Effective design and use of requirements engineering training games." Proc. 7th Int. WS Requirements Eng Edu & Train, 2008.
[31] G. Gabrysiak, M. Guentert, R. Hebig, H. Giese, "Teaching requirements engineering with authentic stakeholders: Towards a scalable course setting." Proc. 1st Int WS Soft Eng Edu Based on Real-World Exp. 2012.
[32] L. Likkel, "Calibrated Peer Review Essays Increase Students Confidence in Assessing Their Own Writing." J College Science Teaching 41(3), 2012, pp. 42-47.
[33] M. Walvoor, M. Haefnagels, D. Gaffin, M. Chumchal, D. Long, "An Analysis of Calibrated Peer Review (CPR) in a Science Lecture Classroom." J College Sci Teach 37(4), 2008, pp. 66-73.
[34] L. Margerum, M. Gulsrud, R. Manlapez, R. Rebong, A. Love, "Application of Calibrated Peer Reivew (CPR) Writing Assignments to Enhance Experiments with an Environmental Chemistry Focus." J Chemical Edu 84(2), 2007, pp. 292.
[35] J. Prichard, "Writing to Learn: An Evaluation of the Calibrated Peer Review™ Program in Two Neuroscience Courses." J Undergraduate Neurosci Edu 4(1), 2005.
[36] B. Furman, W. Robinson, "Improving Engineering Report Writing with Calibrated Peer Review™." Proc. 33rd Ann Symp American Soc Eng Edu, 2003.
[37] P. Black, D. William, "Inside the Black Box: Raising Standards through Classroom Assessment." Phi Delta Kappan 92(1), pp. 81-90.
[38] J. Russel, S. van Horne, A. Ward, E. Bettis III, J. Gikonyo, "Variability in students' evaluating processes in peer assessment with calibrated peer review." J Comp Assisted Learning, 2017.
[39] S. Balfour, "Assessing Writing in MOOCs: Automated Eassay Scoring and Calibrated Peer Review™." Research and Practice Assessment 8, 2013, pp. 40-48.
[40] M. Aniche, F. Mulder, F. Hermanns, "Grading 600+ Students: A Case Study on Peer and Self Grading." Proc 43rd Int. Conf Soft Eng: Soft Eng Edu & Train, 2021.
[41] M. Daun, A. Salmon, T. Weyer, K. Pohl, B. Tenbergen, "Project-based learning with examples from industry in university courses: An experience report from an undergraduate requirements engineering course." Proc. IEEE 29th Int. Conf. Soft Eng Edu & Train, 2016.
[42] NY State, Office of the Governor Press release from 3/11/20, accessed 6/2/21, available at: https://www.governor.ny.gov/news/during-novel-coronavirus-briefing-governor-cuomo-announces-new-york-state-will-contract-28
[43] US Federal Code of Regulation 34 CFR Part 99, "Family Educational Rights and Privacy," available at https://www.law.cornell.edu/cfr/text/34/part-99, accessed 6/2/21.
[44] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd Ed.. Lawrence Erlbaum Assoc, 1988.